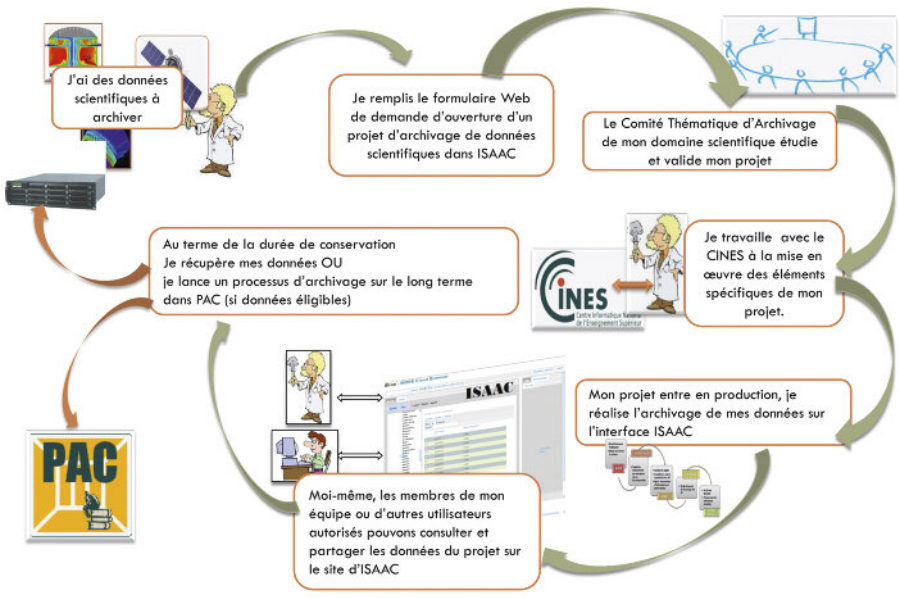
Les données scientifiques, qu'elles soient issues de différents capteurs ou le résultat de simulations, voient leurs volumes et leur diversité croître de manière exponentielle *(big data)* et requièrent des infrastructures et des solutions permettant leur traitement (acquisition, analyse, visualisation, préservation) et leur partage. Les deux missions stratégiques nationales du Cines, calcul intensif et archivage pérenne, le positionnent naturellement pour fournir ce type de services.

Archivage intermédiaire de données scientifiques au Cines

Le Cines participe au projet Eudat (*EUropean DATa*)[1] qui a pour objectif la mise en place d'une infrastructure collaborative destinée à la conservation et au partage des données provenant des communautés européennes de chercheurs. Il est le seul centre de calcul français constituant un nœud de l'infrastructure Eudat. Le service de base offert en termes de pérennisation est le stockage sécurisé des fichiers ou « préservation du train de bits » (qui consiste à s'assurer que la succession de 0 et de 1 « incrustée » dans le média informatique n'est pas altérée au fil du temps) avec l'ajout d'un identifiant pérenne. Son intégrité et son accessibilité au travers d'un tel identifiant sont ainsi assurées contrairement à sa lisibilité (i.e. capacité à ouvrir un fichier « ancien » sur un système informatique « récent » et à en lire le contenu) et son intelligibilité au cours des années (i.e. capacité à comprendre le contenu d'un document archivé malgré le temps qui passe – sans la pierre de Rosette, les hiéroglyphes seraient lisibles mais très probablement inintelligibles), que seul un service d'archivage saura garantir.



Différentes étapes de la vie d'un projet d'archivage dans Isaac.

ARCHIVER QUOI ?

Lors des phases initiales du cycle de vie des données, les scientifiques qui les produisent ont des difficultés pour déterminer celles qui nécessiteront un archivage définitif. La mise en place d'un projet d'archivage pérenne sur la plateforme d'archivage au Cines (PAC) nécessitant un travail conséquent, l'incertitude sur les besoins constitue un frein à l'archivage des données scientifiques.

Afin de mieux comprendre leurs besoins, les équipes du Cines ont mené une enquête auprès d'environ 150 laboratoires de recherche français qui a fait apparaître les points suivants :

- il existe une volonté de conservation et d'accès aux données pendant au moins 3 à 5 ans ;
- le type de données correspond souvent à des résultats de calcul et d'observation, des codes sources ;
- habituellement les données sont conservées dans les laboratoires dans des formats extrêmement diversifiés ;
- l'archivage des fichiers explicatifs, et des métadonnées embarquées dans les fichiers, doit être pris en compte ;
- il n'existe pas de jeux ou de standard de métadonnées par thématique de recherche ;
- un partage des données dans un cercle restreint et la gestion des droits sur ce partage ont été décrits comme des nécessités ;
- les besoins en termes de volume de stockage sont significatifs : on parle de un à plusieurs dizaines de téraoctets par projet.

UNE OFFRE COMPLÉMENTAIRE D'ARCHIVAGE INTERMÉDIAIRE

Pour offrir une solution en adéquation avec les besoins, le Cines a mis en place un service « d'archives intermédiaires de données scientifiques » pour des communautés d'utilisateurs structurées. Ce service, nommé Isaac (Information scientifique archivée au Cines)[2], correspond à un stockage sécurisé des données, comportant obligatoirement un jeu minimal de métadonnées descriptives associées, afin d'en faciliter la recherche et la compré-

hension, pour une période déterminée (3 à 5 ans), et offre une possibilité d'accès partagé à ces données pour la communauté. Au terme de ce laps de temps, selon la décision de leur producteur, les données sélectionnées seront soit versées dans un système d'archivage définitif, soit restituées à l'utilisateur pour destruction ou simple stockage dans son laboratoire.

Bien évidemment, ce service ne peut être fourni que pour une communauté d'utilisateurs « structurée » partageant les même formats de données (formats d'échange standards qui sont utilisés pour l'archive) et désireuse de mettre en œuvre une démarche combinant diffusion et conservation à long terme. Ce n'est pas une solution proposée pour du stockage prolongé des données, mais bien une approche pour faciliter la préparation d'une archive pérenne.

DES FACTEURS DE SUCCÈS

Lors de la mise en place de la plateforme Isaac et des différents services offerts au travers du nœud eudat@cines, le Cines a identifié plusieurs facteurs qui peuvent assurer le succès des projets d'archivage.

• ***Le besoin d'informer sur l'archivage*** : il est nécessaire de communiquer auprès des utilisateurs potentiels sur la plus-value de l'archivage électronique par rapport à un service de « sauvegarde sécurisée », couramment offert par la plupart des directions des systèmes d'information. Ceci doit permettre d'identifier les données nécessitant un archivage et de justifier le travail nécessaire à sa mise en œuvre.

• ***Une relation de confiance*** : la plupart des enquêtes effectuées auprès des scientifiques indique que confier ses données à un tiers nécessite la mise en place d'une relation de confiance. Pour ceci, le Cines met en place des conventions précisant clairement les notions de propriété et de responsabilité. Il continue sa démarche qualité s'appuyant sur les certifications adéquates, par exemple le *Data Seal of Approval* pour la plateforme Isaac.

• ***L'intégration de l'archivage dans la chaîne de gestion des données*** : les processus liés à l'archivage des données sont souvent considérés comme un surcroît de travail. Un des objectifs du Cines est de faciliter l'intégration de cette préoccupation et des actions induites dans les processus de gestion des données afin, par exemple, d'automatiser certains traitements ou de capturer certaines métadonnées dès la création des données.

• ***La migration vers l'archivage pérenne*** : un des risques dont les utilisateurs doivent être conscients est le manque de caractérisation des données ou du format des fichiers lors de leur versement dans l'archive intermédiaire et, donc, une perte possible de la connaissance du contexte de production de l'information lorsqu'on se décide ultérieurement à les verser dans l'archive définitive. Cette recherche d'équilibre entre la simplicité de versement initial et la capacité à la pérennisation doit être évaluée avec les utilisateurs.

• ***La volumétrie des données*** : la volumétrie des données scientifiques à archiver pose plusieurs problèmes techniques qu'il faut résoudre pour garantir une qualité de service lors des dépôts ou des accès.

Marion Massol
Responsable du département Archivage et diffusion, Cines
massol@cines.fr

Stéphane Coutin
Chef de projet, Cines
coutin@cines.fr

[1] www.eudat.eu
[2] www.cines.fr/spip.php?rubrique369

● ● ● QU'ES AQUÒ ?

Retour sur quelques termes croisés dans le dossier de ce numéro d'*Arabesques*.

Big data
Consiste à intégrer, synchroniser, traiter et valoriser de très grands volumes de données informatiques, extrêmement variées et de différentes natures. Ce terme s'est largement répandu en 2012. À l'opposé, les *small data* concernent des données qui peuvent être gérées directement par les individus.

Data mining
Traduit par « exploration de données », « fouille de données » ou encore « extraction de connaissances à partir de données » a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données par des méthodes automatiques ou semi-automatiques.

Data warehouse
Traduit par « entrepôt de données », désigne une base de données utilisée pour collecter, ordonner, et stocker des informations provenant de base de données opérationnelles et fournir ainsi une aide à la décision pour un établissement ou une entreprise. L'Amue, par exemple, a mis en place un entrepôt de données pour faciliter le développement d'outils de pilotage pour les établissements d'enseignement supérieur.

Digital Object Identifier (DOI)
Littéralement « identifiant d'objet numérique », mécanisme d'identification de ressources numériques, comme un film, un rapport, des articles scientifiques, etc. Le DOI d'un document permet notamment une identification pérenne de celui-ci et, par exemple, de retrouver l'emplacement d'un document en ligne si son URL a changé. Les DOI facilitent l'utilisation des bases de données bibliographiques, des logiciels de gestion bibliographique et permettent de produire des citations plus fiables et plus pérennes.

Repository
Traduit par « dépôt », « référentiel » (ou même « répositoire »), désigne un stockage centralisé et organisé de données. Ce peut être une ou plusieurs bases de données où les fichiers sont localisés en vue de leur distribution sur le réseau ou bien un endroit directement accessible aux utilisateurs.

Source : Définitions principalement extraites de Wikipédia.